Introduction

Núria Bel, Benoît Sagot

Introduction

Ressources linguistiques libres
Free Language Resources

1. Language Resources

During the last two decades, the growing importance of empirical approaches to Natural Language Processing has significantly changed the design, construction, operation and evaluation of the so-called Language Resources. In particular, the demand for raw and annotated corpora, and to a less extend lexical data, is even more massive than before, as such resources have become major sources of linguistic information. Unfortunately, the compilation and development of Language Resources is a long, expensive and highly error-prone task, regardless of the approaches used. This is because any Language Resource, if it is large-coverage, constitutes a huge set of complex data. Moreover, most Language Resources are specific to one language, sometimes several languages, and cannot be easily adapted to other ones.

This is one of the reasons for the increasing importance of resources which are freely available: authors and owners, while retaining respectively the authorship and intellectual property, allow for the exploitation, transformation and redistribution of such resources, provided the conditions detailed in the associated license are respected. [1]

Commonly demanded Language Resources, which can be textual, oral or multimodal, include:

– written or oral corpora, raw or, more interestingly, enriched with annotations of several kinds, which can range from morphosyntactic labels to prosodic information, from dependency graphs to discourse structure markup;

– lexical resources with information about one or more levels of linguistic analysis, and also grammars or other forms of modeling different types of language structures — the boundary between lexicon and grammar depends often on the approach (cf. TAG, lexicon-grammar, etc.);

– complex resources, combining an annotated corpus and an associated lexical database or combining information from different levels of linguistic analysis.

1. Note that freely available resources do not include freely queriable resources, such as online dictionaries or corpora that are free to use online on a small scale but impossible to download, and even less so to modify, reuse or redistribute.

TAL. Volume 52 – n° 3/2011, pages 7 à 16

Language Resources can involve one or several genres (journalistic, literacy, technical, user-generated...). The formal quality of the underlying texts or lexical entries can vary as well, from edited texts to SMSs or spontaneous oral data. Moreover, linguistic resources can be monolingual or multilingual. In the latter case, the extent to which the data in the different languages involved do or do not match may vary. The two most frequent situations are parallel corpora — the same data in at least two languages, with at least document-level alignment — and comparable corpora — data from a same genre and domain in at least two languages.

Free availability is the best way for Language Resource developers and users to address the following decisive issues:

– visibility: this is a precondition to both immediate and long-term availability, as an available but non-visible resource will not be found nor used; this issue is too often underrated, despite many initiatives;

– immediate availability: this is a condition for a resource to be cited, used, and thereby improved; it is also necessary for guaranteeing the replicability of associated scientific results;

– long-term availability: the information contained within freely available resources is unlikely to be lost, even when their original developers stop working on them, if such resources can be redistributed by others, in their original form, in a modified form, or integrated within other resources.

Many Language Resources, for example because they are not freely available, not visible enough or not well maintained, are not used — if not recreated anew by others —, even though they could prove very useful. One of the most striking illustrations of how influential and decisive a Language Resource can be when it is freely distributed is the case of the Princeton WordNet (PWN), which had and still has a decisive influence on many and varied applications of Language Technologies for English as well as for other languages.

Bearing these issues in mind, the goal of this special issue is to focus on the full range of scientific issues related to Language Resources, including:

– modeling of the language data that constitute the resource (both linguistically motivated formal frameworks, representation of linguistic information in the form of structured and/or quantitative data, standards for Language Resources...);

– methodologies for the development of Language Resources, either by hand, completely automatic or hybrid (semi-supervised techniques, interlingual transfer methods, use of pre-annotation tools, validation/correction interfaces, etc.), approaches promoting linguistic relevance while minimizing the "human cost;"

– methods for the validation and evaluation of Language Resources including extrinsic evaluation within NLP systems and for experimental linguistic purposes.

Not all these topics are covered in this special issue, as we shall see in more detail in Section 4. Still, it brings together papers that describe various aspect of the design, development and use of many different kinds of Language Resources. But before

going into more detail about the contents of this volume, we thought it might be helpful to sketch the scientific landscape surrounding Language Resources (Section 2) and the international context within which they are developed (Section 3).

2. The role of Language Resources in computational linguistics

Since the emergence of Natural Language Processing as a research area, which almost dates back to that of computer science, the center of gravity of this field has moved from rule-based approaches to statistical methods. Although earlier work has been done using statistical models, the key period during which that shift reshaped the NLP community is the late 80's and the 90's. Starting from speech recognition, the influence of statistical approaches to NLP quickly percolated towards areas such as part-of-speech tagging, parsing or machine translation and eventually also affected the creation of Language Resources. Such approaches have brought increased performance together with consistent evaluation schemes, and improved coverage and robustness.

One of the most often claimed advantages of statistical methods is that they are based on the automatic induction of linguistic knowledge. This is often contrasted with the tedious, costly and error-prone construction of language descriptions, as required by rule-based methods. Indeed, rule-based methods rely on explicit and often handcrafted linguistic knowledge, such as lexicons and grammars, as required for performing computational linguistic tasks.

However, this supposedly lower cost of statistical approaches is only apparent. Indeed, no less linguistic expertise is required within such approaches as compared with more symbolic approaches, simply because linguistic information is encoded, although now in form of annotations and as part of the data on which such systems are trained. Let us take (syntactic) parsing as an example. It is a fact that a large-scale rule-based handcrafted grammar and the associated morphological and syntatic lexicon encode in a formalized way a huge amount of linguistic knowledge, and their development is therefore highly time-consuming. However, it is no less the case for the syntactically annotated corpora, or treebanks on which statistical parsers are trained, which need to be large enough ("there is no data like more data"). Therefore, whatever the type of approach — symbolic, statistical or hybrid —, Language Resources play a key role, and constitute a major bottleneck when they are not available.

What is more, recent work in several fields of Natural Language Processing have shown that leveraging both annotated corpora and earlier types of resources such as lexicons, thus building hybrid systems, leads to improved systems. This is true, for example, for part-of-speech tagging (Denis and Sagot, 2009),multi-word unit detection, named-entity recognition (Cohen and Sarawagi, 2004; Kazama and Torisawa, 2007), or parsing (Goldberg *et al.*, 2009).

However, the use of Language Resources for such tasks is impacted by an important issue, strongly related to the cost of Language Resource development, that of domain adaptation: resources developed for one particular language variety, e.g., a specific genre or domain (often journalistic), typically lead to performance drops when used for processing data from other genres or domains (e.g., technical domains or user-generated content, to name but a few). This issue is even more crucial for statistical systems trained on annotated data, which are often limited to one genre, as the distribution and even the very nature of the data might change from one domaine to another. Hence the development of a large variety of domain adaptation techniques, often based on semi-supervized or unsupervized approaches. [2]

But this raises one of the major concerns about Language Resources: how can they be evaluated? Intrinsic evaluations (accuracy, coverage) can only be obtained by comparison with a gold standard, which is exactly what most Language Resources aim at being. Moreover, many Language Resources rely on a particular formalization, i.e., interpretation of the data, which affects the very notion of correctness. Another way to evaluate Language Resources is to use task-based evaluation schemes: in that case, resources can be compared with respect to the performance of evaluable tools that rely on them. However, this requires interoperability between the Language Resources in hand, as they have to be fit in an otherwise identical applicative context.

Independently of the way Language Resources are used, i.e., irrespective of the fact they are used as training, evaluation, input or external data for symbolic, statistical or hybrid systems, several techniques have been deployed for developing Language Resources, in order to reduce the development cost while preserving or even improving their quality. Indeed, although some Language Resources have been developed for years or decades (e.g., the Princeton WordNet, FrameNet, on-Grammar tables...), semi-supervized or unsupervized techniques are often used, either for creating preliminary versions of the resources (that can be validated and corrected manually if required), or for acquiring linguistic information. In some settings, such techniques offer a cheaper way to large coverage, better consistency and/or even better accuracy. In the recent years, another approach has been increasingly used, namely the collaborative development of Language Resources. It is a way to lower resource development cost while sticking to a (mostly) manual setting. Collaborative resource development can take various forms, such as serious games, direct resource development (e.g., wiki resources) or crowdsourcing. [3]

Despite all these efforts, large-scale Language Resources remain costly to develop, and therefore relatively rare, apart for a few widely-studied languages, and above all English. The majority of languages still lack Language Resources, especially free ones. For example, the development of complex resources such as semantically

2. See for instance the Domain Adaptation workshop at the 2010 ACL conference.

3. Although crowdsourcing *per se* is a promising way to make many people, sometimes many experts, collaborate on a resource development task, the most popular crowdsourcing plateform, the Amazon Mechanical Turk, raises important issues that are discussed in the litterature. See for example Adda *et al.* (2011).

annotated corpora is in its earliest stages for a language such as French. The situation is even worse for less widespread languages (e.g., minority languages), which are less interesting commercially or concern a small amount of computational linguists, and most of them are qualified as less-resourced or resource-scarce. However, there is an increasing effort towards alleviating this so-called linguistic digital divide, in particular with normalization efforts when required, the development of BLARKs (basic Language Resource kits), and applications thereof within NLP tools for such languages.

In order to best take advantage of these efforts, an important issue is that of interoperability between resources. This is not only needed for comparing, evaluating, and even merging Language Resources, in order to increase their coverage, richness or even precision. It is also required for building Language Resources that bring together different levels of linguistic annotation, such as syntactic and semantic information in a lexicon, multi-layer annotated corpora (e.g., prosodic and syntactic information for a corpus or oral transcripts, or named entity, syntax and semantic information in a text corpus). Finally, it is a prerequisite for the emergence of reusable Language Resource management tools. In that regard, standardization is one of the key challenges, but research aimed at encoding Language Resources using ontological techniques and frameworks are also important issues.

A Language Resource can have a large influence on the whole research community if it is easily available, high quality and large scale, especially if it is new of its kind. For example, when the Penn TreeBank was released, the parsing community dedicated tremendous efforts towards the development of parsing systems that would perform well on this treebank, given the standard evaluation metrics. Although this resulted in significant advances in statistical parsing, this had the effect to focus the attention of researchers on constituency parsing for English, which is but one of the many settings of interest for parsing. Since then, following a tradition well established for other languages (e.g., Czech), dependency parsing has retrieved a prominent place. The role of morphological analysis for parsing, which is much more crucial for most languages than for English, has been increasingly investigated. The impact of using other kinds of Language Resources, such as subcategorization or semantic lexicons, has been investigated. Significant efforts have been achieved for developing parsing architectures that allow for parsing data that is not of the same genre and quality than Penn TreeBank texts (i.e, *Wall Street Journal* articles), such as text from specialized domains or user-generated content (social media, forums. . .). Still, the Penn TreeBank remains a reference in the field, and one can definitely say that this resource has reshaped several areas of research such as parsing, and still has a major impact on computational linguistics research.

Another example of a highly influential resource, which we already mentioned, is the Princeton WordNet (PWN), a large-scale lexical semantic database for English whose development was initiated in the 80's at Princeton University by George Miller and a team of linguists and psychologists. This resource, which is freely distributed for both research and commercial purposes, had an enormous impact. Several other

projects were initiated to create wordnets for other languages as well as to extend the PWN with various types of additional information, such as alignments with existing ontologies (SUMO, MILO), information concerning domains, sentiment, and others. The PWN and other wordnets aligned to it have served directly or indirectly as reference resources for virtually all semantic analysis system evaluations that required a preexisting inventory of word senses and lexical semantic database. Of course, the PWN has been criticized, for instance because of its very high sense granularity that is often useless and adds many ambiguities, or because of the fact that it relies on an *a priori* inventory of senses. Indeed, word sense induction techniques are increasingly used, but still compared or integrated with wordnet-based approaches.

3. Institutional context

As mentioned above, the development of Language Resources is costly. Therefore, the institutional and sometimes commercial context have always played a major role in the development of NLP in general and Language Resources in particular.

Still, Natural Language Processing is indebted to the availability of collections of language data that linguists working in lexicography, dialectology and corpus linguistics compiled as empirical evidence on which to base their studies. After the first samples developed by academics, the Brown corpus, as early as in the 60's of the last century, the LOB corpus, etc., the largest collections of digitalized written texts for being processed were compiled thanks to institutional support in what were called "Reference Corpus" of different languages, most of them meant to give support to lexicographers following the Cobuild initiative that was, however, privately funded. The French *Trésor de la Langue Française* (1971-1994), the Spanish Real Academia *Corpus de Referencia del Español Actual* (started in 1997), the *Frequency Dictionary of Contemporary Polish* (1967-1990), the *British National corpus* (1991-1994), the *Spoken Dutch corpus* (1998) are just some examples of these efforts to compile language data. However, they were not conceived for being open, freely distributable resources, one of the reasons could have been the copyrights of the source data. In most cases, they are currently accessible by Web exploitation applications the aim of which is to assist researchers in the study of languages.

Nevertheless, there is early evidence that these resources could have been used for the advancement of language technologies and that their availability would have largely benefited the progress of the field. The development of automatic part of speech tagging statistical methods was based on the Brown Corpus (Jelinek, 1985; Church, 1988; Brill, 1992), and the availability of the Hansards, the proceedings of the Canadian parliament kept in both French and English, was crucial for the first Statistical Machine Translation systems (Brown *et al.*, 1990).

At the same time, a demand for Language Resources to cater text processors with language extended capabilities such as spell-checkers and the success of speech recognition and generation commercial tools were the forces behind the constitution

of other Language Resources, this time protected for commercial interests. It is also worth noting that not all languages were supported by the same commercial demands, what caused different timings in the availability of such tools. Additionally, these first "early" business cases are most probably behind the reluctance of many LR developers to offer the compiled resources as open and for free.

The specificity of the European market, with a number of languages to cover was made clear. The European Commission, in particular the DGXIII, which had actively supported machine translation systems (SYSTRAN and EUROTRA were the first) since the late 70's, intervened in favour of the multilingual Europe with a first R&D funding program called Language Resources and Engineering[4] (1990-1994) with the following objectives: "to create methods, tools and linguistic resources, especially portable software tools, grammars, dictionaries, domain specific terminological collections, as well as large, high-quality corpora and the stimulation of standards work." A number of projects were funded by this program and its successors. NERC, Network of European Reference Corpora (1993-1994) and MULTEXT (1994-1995) and MULTEXT-EAST (1995-1997) were the first of a series of projects that tried to ensure the provision of resources and tools for generic processes and for a broad number of European languages. All project results were to be distributed and be publicly available. Other EU projects followed for offering Language Resources for different applications: PAROLE[5] (1996-1998), SIMPLE[6] (1998-2000) and SpeechDat[7] (1997-1998) were attempts to tackle all the aspects involved in the building of resources: production, standardization, provision and reusability, for text and speech resources. At the same time, the European Language Resources Association (ELRA, founded in 1996) and its distribution agency, ELDA, were set up with the objective of providing support to the market of Language Resources and its stakeholders with a repository of Language Resources and activities such as the International Conference on Language Resources and Evaluation (LREC) that every two years provides an overview of the state-of-the-art in topics related with the availability of Language Resources.

Despite these EC driven initiatives, it took some time for the member states to acknowledge the importance of the field, and only a few created their own specific programs to support their languages. The French Technolangue[8] (2003-2005) is one of these programs. Besides, the Centre National de Ressources Textuelles et Lexicales[9] was set in 2005 as the basis for the distribution of open Language Resources. The Dutch STEVIN (Essential Speech and Language Technology Resources for Dutch) program (2004-2012) has been another of such program for

4. http://cordis.europa.eu/search/index.cfm?fuseaction=prog.document&PG_RCN=177160
5. http://www.elda.org/catalogue/en/text/doc/parole.html
6. http://www.ub.edu/gilcub/SIMPLE/simple.html
7. http://www.speechdat.org
8. http://www.technolangue.net/
9. http://www.cnrtl.fr/

Dutch. STEVIN is funded by the Flemish and the Dutch Governments, both interested in strengthening the economic and cultural position of the Dutch language in the modern ICT-based society. STEVIN also foresaw the use and re-use of the program results: all materials are made available through TST Centrale, the Dutch-Flemish Agency for Dutch Language Resources.

Recently, and again under the funding of EU programs, a number of new initiatives are trying to get attention for supporting Language Resources. The best known are CLARIN, FLaReNet and META-NET. [10] These are three networks that among their objectives share to foster the access to Language Resources for different uses and objective audience. These initiatives show the progressive consolidation of the field and are echoing again the social concern about the need to cover all languages in Europe and in particular the urgency of being able to cater the demand of a very dynamic ICT technology that proposes a wider range of applications that increasingly needs more and richer resources. The solution proposed by the three initiatives is the creation of the infrastructure for the easy discovery and access to available resources.

CLARIN-ERIC (Common Language Resources and Technology Infrastructure) is already a European Research Infrastructure proposing interoperability, federated access to resources and tools, and provisions for the curation, maintenance and accessibility to a broad range of resources and tools for the research in the Humanities.

FLaReNet (Fostering Language Resources Network) has consulted the community about current priorities and get consensus about the strategic objectives that have to drive the field in the years to come. The community agrees that multilingualism has to get further institutional support at all levels, and that Language Technologies are the enabling key for making multilingualism in Europe economically sustainable (Calzolari *et al.*, 2012).

META-NET (Multilingual Technology Alliance Network) is a first step in the direction signalled by FLaReNet and it is currently leading the development of META-SHARE, [11] a distributed platform to host information about available and accessible Language Resources which promotes the actual sharing of resources.

4. Reading roadmap

This special issue devoted to open Language Resources offers a nice overview of the current state-of-the-art in the development of free Language Resources: those that can be inspected, transformed and exploited. The nine papers finally selected by the specific program committee (see below) show samples of resources in three different languages, but, more crucially, give a comprehensive list of the issues that turn out to be important for releasing a free Language Resource. The papers cover topics related to: the recovery of legacy resources (Jacobson and Baude; and Tolone);

10. `http://www.clarin.eu/`, `http://www.flarenet.eu`, `http://www.meta-net.eu`.
11. `http://metashare.dfki.eu`

the maintenance and extension of existing resources (Vincze and Bestgen; Tolone); the production of new resources: how to design them (Péry-Woodley *et al.*; Balvet *et al.*), how to collect and annotate resources using crowdsourcing (Vetere *et al.*) and bootsrapping techniques that profit of machine learning methods (Péry-Woodley *et al.*; Vincze and Bestgen), and also how to describe the resource with appropriate metadata and its derived benefits (Jacobson and Baude); the issues related to the actual open resource distribution in terms of formats (Péry-Woodley *et al.* report on TEI-P5 XML corpus encoding; Chiarcos *et al.* and Vetere *et al.* comment on linking resources to OWL; Balvet *et al.* report on a noun lexicon delivered in the Lexical Markup Format, and Tolone on Le*fff*), IPR issues clearance (Pery-Woodley *et al.*; Eskhol *et al.*); and last but not least, the importance of allowing access for those non-language technologies experts with specific interfaces (Falaise *et al.*).

The resources covered in this issue range from written resources (most papers) to speech resources (Eshkol *et al.*; Jacobson and Baude), from text corpora (Péry-Woodley *et al.*; Falaise *et al.*) to lexica (Balvet *et al.*; Vincze and Bestgen; Tolone) and from nominal aspectual characteristics (Balvet *et al.*), discourse annotation (Péry-Woodley *et al.*) to annotation of emotion and polarity words (Vincze and Bestgen). The importance of the usability of previously existing resources is remarkable. Different resources presented in this issue have been built upon developments of other previously available resources like treebanks. The use of tools such as POS-taggers and parsers play also a crucial role in building new resources. It is also important to note the interest of the community in providing open resources for languages other than English. This issue involves French resources, but also Italian and Spanish.

5. Conclusion

This volume addresses a number of issues concerning the compilation and development of Language Resources. From the reading of the different contributions it is easy to assess the enormous effort that has been invested in the task and not only from the linguistic point of view. Accessibility, documentation and formating into current practices are also key steps for making them widely available and, in fact, used. All in all, their development is an expensive task which will not be undertaken unless there is a sustained institutional support. It is important to realize that, despite of the fact that the lack of coverage can leave different languages and topics uncovered, the Language Technologies industry is limited to the market demands and the return of investment for developing resources for their applications to cover all languages and all domains. Thus, the re-use and re-purposing of existing resources will be one of the crucial factors for getting new products. For this to happen, it is of utmost importance that resources are free and easily available. Furthermore, the contributions of this special issue also show that the interest in Language Resources has also been positively affecting linguistic research. The free availability of resources it is important for getting data to support research and to give light to different and still pending issues in the description of languages.

Acknowledgment

We would like to express our gratitude to the whole scientific committee of the TAL journal, and even more so to the members of the specific scientific committee that was set up for this issue:

– Delphine Bernhard, LiLPa, Université de Strasbourg (France)
– M. Teresa Cabré, IULA, Universitat Pompeu Fabra, Barcelona (Spain)
– Emmanuel Cartier, Université Paris-Nord (France)
– Anne Condamines, CLEE-ERSS, CNRS, Toulouse (France)
– Benoit Crabbé, Alpage, Université Paris-Diderot (France)
– Paul Buitelaar, DERI, National University of Ireland, Galway (Ireland)
– Cécile Fabre, CLEE-ERSS, Université Toulouse-Le Mirail (France)
– Mikel L. Forcada, DLSI, Universitat d'Alacant (Spain)
– Gregory Grefenstette, Exalead (France)
– Eva Hajičová, ÚFAL, Univerzita Karlova, Prague (Czech Republic)
– Nancy Ide, DCR, Vassar College (USA)
– Christine Jacquin, LINA, Université de Nantes (France)
– Marie-Claude L'Homme, OLST, Université de Montréal (Canada)
– Anne Lacheret, MoDyCo, Université Paris-Ouest (France)
– Joseph Mariani, LIMSI, CNRS, Orsay (France)
– Piet Mertens, Subfaculteit Taalkunde, Katholieke Universiteit Leuven (Belgium)
– Asunción Moreno, Universitat Politècnica de Catalunya, Barcelona (Spain)
– Gertjan vcd an Noord, Rijksuniversiteit Groningen (The Netherlands)
– Sophie Rosset, LIMSI, CNRS, Orsay (France)
– Agata Savary, IUT de Blois, Université François-Rabelais, Tours (France)
– Jesse Tseng, CLEE-ERSS, CNRS, Toulouse (France)
– Agnès Tutin, LIDILEM, Université Stendhal-Grenoble 3 (France)
– Piek Vossen, Vrije Universiteit Amsterdam (The Netherlands)

Núria Bel
Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelone, Espagne
nuria.bel@upf.edu

Benoît Sagot
ALPAGE, INRIA Paris-Rocquencourt & Université Paris 7, Paris, France
benoit.sagot@inria.fr